Sequence Tolerance of a Highly Stable Single Domain Antibody: Comparison of Computational and Experimental Profiles

Mark A. Olson,[1] Patricia M. Legler,[2] Dan Zabetakis,[2] Kendrick B. Turner,[2] George P. Anderson[2] and Ellen R. Goldman[2]

[1]Department of Cell Biology and Biochemistry, Molecular and Translational Sciences, USAMRIID, Frederick, Maryland, United States of America

[2]Center for Bio/Molecular Science and Engineering, Naval Research Laboratory, 4555 Overlook Ave SW, Washington, D.C., United States of America

Abstract:

The sequence fitness of a single-domain antibody with unusually high thermal stability is explored by a combined computational and experimental study. Starting with the crystallographic structure, RosettaBackrub simulations were applied to model sequence-structure tolerance profiles and identify key substitution sites. Experimental site-directed mutagenesis was used to produce a panel of mutants and their melting temperatures were determined by thermal denaturation. The results reveal an excess stability margin of approximately 12 °C, a value taken from a decrease in the melting temperature of an electrostatic charge reversal substitution in the CRD3 without a deleterious effect on the binding affinity to the antigen target. Tolerance for disruption of antigen recognition without loss in thermal stability was demonstrated by the introduction of a proline in place of a tyrosine in the CDR2, producing a mutant that eliminated binding. To reconcile the differences between the modeled energies and their relationship to the observed experimental changes in melting temperatures, an approximation was developed by combining a statistical potential with a linearly scaled implicit solvent model to calculate the net contribution from a two-state model of folded and unfolded conformations. The derived computational model improves prediction accuracy and should prove applicable to other designs of antibodies.

Author for correspondence: Mark A. Olson

1. Introduction

Combinatorial optimization of protein sequences to fit a topological fold takes place on comparability landscapes with fitness measured by scaling factors [1]. From an experimental perspective, sequence-structure tolerance is typically probed by site-directed mutagenesis and the landscape is thermal stability, protein fold reorganization and conservation of protein function. A common observation of mutational studies is the remarkable plasticity of protein folds to amino acid changes and how large the sequence envelope is for particular fold families.

A well-known class of protein folds that occupies a superfamily of sequences is the N-terminal domains of both the heavy and light chains of the variable region of antibodies. A popular construct of the heavy chain is single-domain antibody (sdAb) chains derived from camelids. The interest in sdAbs lies in their biotechnological applications that require evaluated thermal stability and reversible folding, as well as high affinity and singularity of molecular recognition [2-7].

Typical sdAbs have melting transition temperatures (T_m) in the range of 60-70 °C [2-5]. An outlier is a llama sdAb specific for *Staphylococcus aureus* enterotoxin B (SEB), which has a reported T_m of 85 °C [2]. The X-ray crystallographic structure of this unusually stable sdAb (designated as A3) reveals an asymmetrical homodimeric assembly of conformers displaying differences in secondary structure geometry and local connectivity [8]. The protein fold topology of each monomer is the common assembly of two β sheets with a β-sandwich arrangement. Structural alignment search with PDB entries show strong fold neighbors with high Z-score matches with other antibody structures.

From a computational perspective, the effect of a mutation on T_m can be calculated from all-atom simulations of thermodynamic folding-unfolding dynamics that govern the heat capacity or melting curve [1,9-11]. Alternatively, an alchemical process of residue mutation can be modeled by using free energy perturbation theory applied to a cycle of the folded and unfolded states to determine changes in free energy of folding stability [12-14]. While both modeling methods are rigorous, they are computationally prohibitive when applied to a large set of mutants for proteins of moderate size or larger. Because of this drawback, algorithms have been developed

with energy functions specifically parameterized to predict differences in the folding free energy due to point mutations. Examples include Site Directed Mutator (SDM) [15], FoldX [16], PoPMuSiC [17], and PreTherMut [18], among others [19,20]. Typically these algorithms are empirically weighted using data obtained from protein engineering experiments and their predicted free-energy changes are correlated to experimental differences in the folding free energy.

Here, this work presents a combined computational and experimental study of identifying residue contributions that govern the unusually high T_m of A3. Our earlier study reported several alanine substitutions and their application to assess comparative modeling methods to predict thermal stability [21]. The computational strategy here is different and is based on RosettaBackrub simulations [22,23] to model the sequence-structure tolerance landscape and generate conformations for a set of amino acid substitutions including non-alanine mutations. Ranking of the mutants and their conformations is attained by the application of the empirical energy function Rosetta and an alternative all-atom statistical potential. From our computational analysis, we apply site-directed mutagenesis to generate a panel of mutants that occupy structural regions identified as possible sequence "hot spots" and to evaluate the accuracy of modeling stability. Each mutant is experimentally characterized by their change in T_m relative to the wild-type A3 monomer and, unlike previous work [21], their binding affinity to the protein target SEB is reported. The latter is important in understanding the sequence threshold of protein function.

To develop a more accurate computational model for the design and selection of A3 mutants, we construct a linear scaling approximation to reconcile the differences between the computed energies for the sequence substitutions and their relationship to the experimentally measured T_m changes. Rather than the more conventional approach of predicting the correlation between an energy function and experimental folding free energies, we examine the often occurring case where the only available experimental data from protein engineering are T_m measurements for selected mutations, as is the case for A3. The empirical nature of constructing a relationship between the T_m and the folding free energy (measured near or below 300 K) is nicely illustrated by the experimental study on the mini-protein Trp-cage with multiple sequence mutations [24]. The general relationship depends on the intrinsic physical properties of each protein under investigation, as well as the experimental conditions of denaturation.

Our modeling results of A3 will show that, while the overall trend of changes in T_m is captured for a subset of residues by combining conformational sampling of the folded state and a statistical potential to score conformations, there a several significant outliers that the routine application of these methods failed to correctly model. We will show that this approach can be improved by a scaling approximation that combines the statistical potential with a weighted generalized Born solvent model to account for dipolar reorganization from the mutations. The approximation is applied to the folded conformations from RosettaBackrub simulations and unfolded conformations determined from temperature-based replica-exchange (T-ReX) dynamics. While the proposed strategy is approximate in its nature as are other common modeling techniques, the method is efficient and provides improvement in the overall correlation of predictions with experiments for this highly thermal stable antibody chain.

2. Computational and experimental methods

2.1 Computational approach

Starting conformations for modeling were taken from the recently reported crystallographic structure of the A3 sdAb determined to a resolution of 2.13 Å (PDB 4TYU) [8]. The structure is a dimeric assembly of two distinctive chains, one denoted as the A conformer and the other the B conformer. Because of the lack of a fully formed disulfide bond in the electron density maps of each chain, the S-S bond was modeled. Sequence tolerance profiles for each conformer were generated by the RosettaBackrub simulation method developed by Smith and Kortenne [22]. The RosettaBackrub protocol consisted of Monte Carlo simulations of flexible backbone and side chain moves in the Rosetta modeling program. The simulations were followed by a combination of simulated annealing and genetic algorithm optimization methods in RosettaBackrub to enrich for low-energy sequences.

For the application here, we selected to model the sequence-fitness profiles for regions F29-M34, P55-Y60, Y98-K104, and M111-Y118, where amino acids were ranked individually for each sequence position by a predicted frequency of tolerance [22]. The generalized Rosetta modeling version [25] was applied and the number of generated conformations was set to 20 for each residue position, the Boltzmann factor set to 0.228, and the fitness score reweighting was

given a value of 0.4 [23]. Sampling convergence from RosettaBackrub was established by noting little change from increasing the number of generated conformations to 50 for a select set of mutations.

Based on the calculated profiles, specific sequence sites were selected for substitution with alternative amino acids and their conformations were generated by the RosettaBackrub simulation method. Simulation parameters were set similar to that of computing the sequence-fitness profiles. Two scoring functions were applied to the generated conformation ensembles and consisted of the Rosetta 3.1 energy function [25] and the statistical potential dDFIRE [26].

For the wild-type (WT) form and selected mutants, short-time T-ReX simulations were carried out using the self-guided Langevin dynamics (SGLD) simulation method to sample conformational space [11,27]. Starting conformations for the mutants were taken from RosettaBackrub simulations as starting decoys for structure refinement and predictions. All decoys plus the corresponding X-ray crystal structure were subjected to energy minimization by the method of steepest descent minimization for 50 steps using the CHARMM22 force field with the CMAP backbone dihedral cross-term extension potential [28]. Solvent effects were modeled using the generalized Born (GBMV2) implicit solvent model [29]. The GBMV2 parameters were set to values of $\beta = -12$ and $P3 = 0.65$ to smooth the energy surface. The hydrophobic cavitation term was modeled by applying the solvent-exposed surface area of the protein solute with a surface tension coefficient set to a value of 0.015 kcal/mol/Å^2.

The SGLD simulations were carried out using the CHARMM22 + CMAP/GBMV2 potential energy function. An integration time step of 2 fs was used for all simulations. SGLD parameters of the friction constant was set to γ of 1 ps^{-1} for all heavy atoms, the guiding factor λ set to a value of 1, and the averaging time t_L was set to 1 ps. Selection of these values was taken from our previous studies of the SGLD model [11,30]. Non-bonded interaction cutoff parameters for electrostatics and vdW terms were set at a radius of 22 Å with a 2-Å potential switching function. Covalent bonds between the heavy atoms and hydrogen atoms were constrained by the SHAKE algorithm [31]. All protein targets during the simulations were unconstrained, freely to reorganize from conformational sampling.

Replica-exchange simulations were performed using the MMTSB [32] utilities and programming libraries for implementing the CHARMM simulation program (version c33b2) [33]. Simulations were carried out using 32 replica clients and frequency of exchanges was set to every 1 ps of simulation. The lower and upper bound temperatures were set at T_{min} = 300 K and T_{max} = 475 K.

2.2 Mutagenesis and characterization methods

Site-directed mutagenesis for the construction of 18 mutants of the A3 sdAb was performed using the QuikChange site-directed mutagenesis kit (Agilent); mutations were confirmed by sequencing (Operon). Protein was expressed and purified from the periplasm as a monomer using a combination of osmotic shock, immobilized metal affinity chromatography, and size exclusion chromatography as described previously [2,4]. The purified protein used for all experiments was in a monomeric form. Protein concentration was determined based on absorbance at 280 nm using a nanodrop 1000 Spectropolarimeter (ThermoFisher). Samples were stored refrigerated in phosphate buffered saline until characterization.

The T_m of each mutant was measured by circular dichroism (CD) using a Jasco J-815 CD Spectropolarimeter equipped with a PTC-423S Peltier for temperature control as previously described [3,4]. The CD measurements were done at least in duplicate, often with several different preparations of the protein. The T_m values estimated from replicate measurements made by CD were all within less than a half degree of each other.

Surface plasmon resonance (SPR) utilizing a ProteOn XPR36 (Bio-Rad), was employed to determine the binding of each mutant to surface immobilized SEB antigen [2,4]. All SPR measurements were done at least in duplicate, often with several different preparations of each mutant as a monomer. The K_D determined by SPR for replicates agreed within at least a factor of three. Other than the mutants at P55, the K_D values were considered to be essentially equivalent to wild-type, as K_D determinations can be problematic when off rates are so slow.

3. Results and discussion

3.1 Structural features of A3

The X-ray crystallographic structure of the llama A3 sdAb is illustrated in Figure 1 and shows an unusual asymmetric assembly of two conformers (denoted as A and B chains) that differ in their secondary structure [8]. The structural differences in the conformers are primarily centered on residues S50-Y60 of linking the β-strands of the highly variable complementarity determining region CDR2. The dimeric interface consists of residues 113-121 of the CDR3 loop as well as E44 and R45 of a variable β-hairpin, and Y98 of the framework.

Because of the rare conformational assembly of A3 relative to other reported crystallographic structures of antibodies and the high T_m, a structural alignment search of the A and B chains was performed by the Dali algorithm [34] to detect potential regions for residue substitutions. As anticipated, the search finds greater than 1000 structural neighbors with high Z-scores and sequence identities that span the range from 77% to a low of 6%. Shown in Figures 1b-c are the high Z-score ranking sdAb hits of PDB entries 3stb and 1i3v with sequence identities of roughly 77% and 66%, respectively. Among the multiple neighbors observed from Dali, the pairwise alignments show structural variability in the length and placement of the α-helix for residues 26-32, β-turn region centered at P55, and the loop region of roughly 98-110.

To explore the relationship between the Dali search results and sequence fitness of structural regions to amino acid substitutions, we calculated the sequence tolerance profiles that contribute to fold stability of the conformers using the RosettaBackrub simulation method [22,23]. Figure 2 illustrates profiles for selected regions F29-M34, P55-Y60, and Y98-K104. The profiles report the ranking of amino acids for each sequence position by a predicted frequency of site population. Wild-type residues are shown in red and the dashed line indicates a cutoff of picking the top 5 amino acid choices at each position.

The computed profiles illustrate variation between the A and B conformers at specific sequence sites where structural differences are most evident. One example is the Y59-Y60 segment, where the aromatic rings are less favorable for the A-chain conformer and are located in an unstructured topology, whereas a β-strand is found for the B-chain. Despite the structural

differences between the conformers, WT residues with high favorable ranking are observed for many sites, suggesting strong sequence-fold comparability fitness. The helix at residue 29 is found in the B-chain conformer, while this secondary structure element is missing in the A-chain, yet the WT phenylalanine is nearly equivalent in ranking of both chains. Conversely, the WT P55 is determined to be of low frequency for both conformers among sequence families.

3.2 Experimental site-directed mutagenesis

To test the computed tolerance profiles and the effect of structural variability observed from the Dali search results on fold conservation, a panel of substitutions for experimental site-directed mutagenesis were selected and consisted of 18 mutants (listed in Table 1). Thermal denaturation curves were determined for each protein mutant (as a monomer) along with their binding kinetics to SEB to ensure proper folding as determined by their ability to recognize the antigen. Representative CD and SPR data are shown in Figure 3. With the exception of the Y59P, all the mutants retained their binding function and most recognized SEB with near wild-type affinities. The mutations at P55 and Y59 may affect the conformation of CDR2 (see Figure 1), which was previously shown to provide critical interactions between A3 and SEB [3]. The melting transition temperatures were compared to the WT form.

It is of interest to place the experimental measured results in the context of the computed profiles of Figure 2 and apply the ideas of protein fitness [1]. Among the mutants, an asymptotic margin or threshold robustness of excess configurational stability can be estimated from the substitutions of which retained the functional property of native-like binding to the antigen SEB. The most notable outcome is the electrostatic charge-reversal D102R in CDR3, which showed a significant drop of 12 °C in the T_m compared to the WT form and demonstrated a native-like K_D (Table 1). The D102R T_m brings the stability excess near the upper bound that is typical of other sdAbs. Additional mutants with excess stability of ~ 10 °C include R70A and Y98A, both positioned in FR3. In contrast to D102R, elimination of the atomic charge with D102A showed a negligible ΔT_m and binds with equal affinity to the antigen as observed with the WT form. Figure 2 reports the WT D102 placed in the high-frequency selection for both the A and B chains, and the D102A mutant was correctly predicted to retain fold stability, whereas D102R was predicted to be weaker.

In contrast to D102R, the limit of functional viability from sequence tolerance is observed by the introduction of a proline in place of a tyrosine (Y59P) in CDR2, producing a mutant where binding was eliminated while thermal stability remained near the wild-type form. The interplay of stability and function of the CDR2 is further established by the P55S mutant where the K_D is decreased and the loss in stability is negligible. Because of the low sequence consensus of proline in all the computed profiles, the substitution P55S and the proline introduction Y59P are difficult to correctly predict from RosettaBackrub.

The mutant Y98A with an experimental measurement of $-\Delta T_m \sim 3$ °C suggests the tolerance profile for the B-chain to be more accurate in describing configurational stability than that of the A-chain, which is thought to be a kinetically trapped chain in the dimeric form [8]. A similar observation can be made for mutants F29A, F29L and K104G.

3.3 Comparison of modeling and experimental data

As an initial step in constructing an empirical relationship between the experimental T_m and a predictive model at the molecular level, RosettaBackrub simulations were applied to calculate the effects of residue substitutions on changes in the Rosetta scoring of conformers (denoted in general by a free energy change ΔG). Figure 4a reports the correlation of using both conformers A and B as input structures to calculate conformational ensembles for each mutation. While it is important to note that the RosettaBackrub calculations only model the folded conformations and do not reproduce the thermodynamic coexistence between the folded and unfolded states as in the experimental T_m, the correlation is nevertheless of general interest to test whether simple models can detect significant outliners in changes of fold stability. The computed Rosetta score is an average over the ensemble and is relative to the WT form taken from the crystallographic assembly. The calculated correlation coefficient of Fig. 4a for evaluating 18 mutants using either the A or B conformer is only $r = \sim 0.2$.

Given the poor performance of approximating the observed experimental changes, the conformational ensembles that were generated by RosettaBackrub for each mutation were rescored using the statistical potential dDFIRE. The results are shown in Fig. 4b and report a correlation of $r = 0.1$ for the A conformer and 0.5 for the B conformer. While calculations for

the B conformer yielded a slight improvement, there are several significant outliers that unfavorably deter an accurate correlation (e.g., F29A, D102R, and K104G).

Rather than using the ensemble of conformations in computing the relative differences, the top scoring conformations determined by dDFIRE were only analyzed. Figure 3c shows the correlation $r = \sim 0$ for the A conformer and $r = 0.3$ for the B conformer. For additional comparison purposes, application of the SDM server [15] was applied to both conformers. SDM is a two-state model calculation and uses a statistical potential energy function that incorporates environment-specific amino-acid substitution frequencies within homologous protein families to calculate a stability score. Figure 4d reports the SDM results and their relationship to ΔT_m. We find a correlation for the A conformer of $r = 0.1$ and $r = 0.4$ for the B conformer.

Selected conformations for several mutants and their reorganization from the WT B-chain conformer predicted by RosettaBackrub simulations are highlighted in Figure 5. The structure for F29A shows distortion in the backbone conformation from the WT form and disrupts the α-helix formation for many of the ensemble. This modeling result combined with sequence-tolerance profiles given in Fig. 2 suggests the helix is not a determinant of stability, but rather poor scoring of conformations. Noticeable backbone displacements are likewise observed for M111A, and to a lesser extent, D102R. Overall, examination of the predicted structures proves difficult to resolve the differences between simulations and experiments.

3.4 Temperature-based conformational sampling

To assess whether the two conformers from the crystallographic assembly need to undergo relaxation as a form of cooperative transition from complex formation prior to RosettaBackrub simulations and which chain in the assembly is more favorable, we applied all-atom T-ReX at a temperature range of 300-475 K, starting with a 50-50 mixture of the A and B conformers. The mixture was selected to allow exchanges among the sampled conformations on the energy landscape to determine which conformer is more populated. Figure 6a reports the simulation results of a probability-density landscape scored by the dDFIRE potential energy. From culling 20,000 conformations at 300 K, we find structures labeled as B-chain to make up approximately 80% of the conformational ensemble that funnels to the lower-temperature cluster, and thus appears to be more favorable. The energy landscape shows two major basins for the B-chain

conformer and the A-chain conformer is less densely populated with excursions covering one of the B-chain basins (Fig. 6a-b).

The top-rank ordered conformations extracted from the three major basins are illustrated in Fig. 6c. Transitions are observed among the peripheral secondary-structure elements of the core β-sheet region, while interestingly the F29 α-helix remains mostly intact in the generated ensemble exchanged to the replica client at 300 K. Using the top-rank ordered A and B conformers from dDFIRE as input to RosettaBackrub, Fig. 6d shows the correlation between predictions and experiments. We find no improvement in the quality of the correlation from structural relaxation.

To better understand how to improve on empirical predictive models, T-ReX simulations were used to check if a physics-based approach can capture the correct trend in thermal unfolding for several mutants in comparison to the WT form. Figure 7 shows the simulation results for mutants F29A, Y98A and D102R, and their comparison with the WT. Presented are the two-dimensional probability-density profiles for the four A3 sequence-structure variants. Unlike the relative scoring of the RosettaBackrub conformations (either Rosetta or dDFIRE scoring), the simulations correctly show little difference between F29A and the WT form, which is consistent with experiments showing a ΔT_m of ~ 1 °C. Significant unraveling is seen for Y98A and D102R with excursions to large RMSD values, while the WT and F29A populated mostly native and near-native RMSD basins. This distinction can be attributed in part to the realistic treatment of electrostatic solvent effects in the GBMV2 model.

3.5 Empirical scaling approximation

To resolve the differences between the dDFIRE scoring of conformations generated from RosettaBackrub and their relationship to the observed experimental changes in melting temperatures, a linear-scaling model was developed by introducing an all-atom solvent free-energy term and taking into account the unfolded state. The former is the application of the GBMV2 implicit solvent model used in sampling conformational space of the crystallographic structure of A3 shown in Figures 6-7. Our linear scaling model is given by the following net sum

$$\Delta\Delta G_{\text{fold-unfold}} = \Delta\Delta G^{\text{dDFIRE}}_{\text{fold-unfold}} + 1/\varepsilon_{\text{f}}\, \Delta G^{\text{GB}}_{\text{fold}} - 1/\varepsilon_{\text{u}}\, \Delta G^{\text{GB}}_{\text{unfold}}, \qquad (3.5.1)$$

where $\Delta\Delta G^{\text{dDFIRE}}_{\text{fold-unfold}}$ is energy difference between a mutant conformation and the wild-type structure computed for both the folded and unfolded forms using the dDFIRE potential function. The term $\Delta G^{\text{GB}}_{\text{fold}}$ is the GBMV2 solvent free energy for the folded conformation computed as the difference between the mutant and wild-type form, and similarly, $\Delta G^{\text{GB}}_{\text{unfold}}$ is for the unfolded conformation.

The terms ε_{f} and ε_{u} in equation (3.5.1) are scaling parameters determined from a linear fit to the experimentally obtained ΔT_{m} for each mutant. To model higher resolution, the scaling parameters can be formulated to account for local environment-specific effects along the protein chain, amino acid type and substitution; namely, $\varepsilon_{\text{f}}(x_i, \alpha \rightarrow \beta)$, where x_i is the spatial location of residue i of type α to be mutated to type β. For modeling the wild-type unfolded state, conformations were generated by T-ReX/SGLD at high temperatures using the CHARMM22/GBMV2 force field. For mutations and their values of $\Delta G^{\text{GB}}_{\text{unfold}}$, side-chains were replaced in the ensemble of unfolded conformations by using the SCWRL modeling program [34] and were subjected to energy minimization using the same force field.

Given the better outcome of using the B-chain conformer in the RosettaBackrub simulations and the T-ReX calculations, we applied equation (3.5.1) to this conformer and computed an ensemble average. Figure 8 reports the linear-scaling model with ε_{f} fitted to ~20 and $\varepsilon_{\text{u}} \sim 10$ for 16 mutants selected from the dataset of 18. The remaining two mutants (S25I and R70A) plus a published third mutant [3] will be used to test "blind" predictions of the model. We find the correlation coefficient to be $r = \sim 0.9$, a significant improvement from the routine application of simulations of modeling only the folded state, scored by either of the two empirical potentials shown in Figures 4 and 6. We should note that merely including the unfolded state scored by dDFIRE is not adequate to improve the correlation, particularly that of D102R and K104G.

The reweighting of the solvent terms in predictive models is not a new idea and has been the subject of many investigations of continuum models for computing solvent energies of

protein structures [36-40]. For the ideal model, $\varepsilon_f = \varepsilon_u = 1$, which describes a condition suitable for modeling substitutions that do not alter the ionization or induce significant conformational changes of surface polarity. For the work presented here, the high value of ε_f reflects structural changes not adequately captured when conformations generated by the RosettaBackrub simulations are introduced to the GBMV2 reaction field for mutations that affect dipolar reorganization. Likewise, ε_u accounts for the lack of an extensive conformational ensemble generated for the mutants.

From the plot of Figure 8, achieving good universality in the fitted parameters appears noticeably weak for several single-point mutants, particularly the charge deletion K104G. This mutant disrupts a salt-bridge of the protein fold and conceivably ion-pair interactions require ε_f to be explicit on chain position and the type of mutation as modeled by $\varepsilon_f(x_i, \alpha \rightarrow \beta)$. To test this idea, we modeled the ion-pair breaking mutant R70A, which was not included in fitting ε_f from the 16-mutation dataset. Using the initial $\varepsilon_f = 20$ and $\varepsilon_u = 10$, the predicted $-\Delta T_m$ for R70A is ~2 °C, while the experimental value is 11 °C. Similarly for K104G, predictions give $-\Delta T_m$ of 3 °C, whereas the reported experimental value is 7 °C. Given the underperformance for these two ion-pair breaking mutants, increasing the contribution of ΔG_{fold}^{GB} by setting $\varepsilon_f = 12$ improves $-\Delta T_m$ predictions for R70A ~ 10 °C and K104G ~ 8 °C. While the value of ε_f appears to be arbitrary, it was selected to give nearly equal weight between the folded and unfolded states for solvent polarization of charge deletion, which is typical of the solvent balance in T-ReX simulations, even though the scaling of both terms is > 1 from the lack of exhaustive conformational sampling.

To further examine the fitted scaling model, we calculated two additional mutants not included in the empirical fit: S25I and T28I. The experimental $-\Delta T_m$ for each mutant is ~ 0 °C for S25I and −2°C for T28I [4], while the predicted values from the scaling model are respectively, ~ 0 °C and −1 °C. The GB solvent terms play different roles in stabilization of the two mutants; for S25I, $\Delta\Delta G_{fold\text{-}unfold}^{dDFIRE} - \Delta G_{fold}^{GB} \sim 0$, while for T28I, the difference favors enhanced stabilization.

When including all known 19 mutations of A3 (Table 1 plus T28I) and using the initial fitted scaling parameters, $r = 0.83$. From modeling this dataset, universality in ε_f among vastly different proteins is likely difficult given the empirical nature of the scaling constant and how it reflects structural characteristics unique to each protein (e.g., salt bridges, hydrophobic packing, etc.). While our goal was not to provide an extensive benchmark, the scaling values obtained for A3 are expected to provide a good starting point for modeling other sdAbs in capturing high rank-order changes in thermal stability from sequence fitness in terms of ΔT_m.

4. Conclusions

The sequence fitness of the sdAb A3 with unusually high thermal stability was investigated by a combined computational and experimental study. An alternative computational strategy was explored for modeling the effect of amino acid substitutions and their correlation to changes in melting temperatures. Our approach is the application of the RosettaBackrub simulation method to model sequence substitutions and their conformational changes from the native structure. For modeling the unfolded state, temperature-based replica-exchange dynamics simulations were applied to generate conformational ensembles. To extend the calculations for a more accurate reproduction of the observed experimental changes in melting temperatures, a linear-scaling model approximation was developed by combining the statistical potential dDFIRE with a scaled generalized Born solvent model to calculate the net contribution from the two-state model of folded and unfolded conformations. While the scaling approximation is simplistic in its approach, good improvement was gained in the rank order of sequence substitutions of a highly thermal stable sdAb.

Acknowledgements

This work was funded by the U.S. Defense Threat Reduction Agency JSTO award CBCALL12-LS6-2-0036 (ERG). Funding support for KT was from the American Society for Engineering Education Postdoctoral Fellowship. The opinions or assertions contained herein belong to the authors and are not necessarily the official views of the U.S. Army, U.S. Navy or the U.S. Department of Defense.

References

1. Tokuriki N, Tawfik DS. 2009 Stability effects of mutations and protein evolvability. Curr Opin Struct Biol 19: 596-604.

2. Graef RR, Anderson GP, Doyle KA, Zabetakis D, Sutton FN, et al. 2011 Isolation of a highly thermal stable lama single domain antibody specific for Staphylococcus aureus enterotoxin B. BMC Biotechnology 11.

3. Zabetakis D, Anderson GP, Bayya N, Goldman ER. 2013 Contributions of the complementarity determining regions to the thermal stability of a single-domain antibody. PLoS ONE 8(10): e77678. doi:10.1371/journal.pone.0077678.

4. Turner KB, Zabetakis D, Goldman ER, Anderson GP. 2014 Enhanced stabilization of a stable single domain antibody for SEB toxin by random mutagenesis and stringent selection. Protein Engineering Design & Selection 27: 89–95.

5. Turner KB, Zabetakis D, Legler P, Goldman ER, Anderson GP. 2014 Isolation and epitope mapping of staphylococcal enterotoxin B single-domain antibodies. Sensors (Basel) 14: 10846–10863.

6. Muyldermans S. 2013 Nanobodies: natural single-domain antibodies. Annual Review of Biochemistry 82: 775–797.

7. Eyer L, Hruska K. 2012 Single-domain antibody fragments derived from heavy-chain antibodies: A review. Veterinarni Medicina 57: 439–513.

8. George J, Compton JR, Leary DH, Olson MA, Legler PM. 2014 Structural and mutational analysis of a monomeric and dimeric form of a single domain antibody with implications for protein misfolding. Proteins: Structure, Function and Bioinformatics 82: 3101–3116.

9. In-Chul Y, Lee MS, Olson MA. 2008 Calculation of protein heat capacity from replica-exchange molecular dynamics simulations with different implicit solvent models. Journal of Physical Chemistry B 112: 15064–15073.

10. Lee MS, Olson MA. 2011 Comparison of two adaptive temperature-based replica exchange methods applied to a sharp phase transition of protein unfolding-folding. Journal of Chemical Physics 134: 2444111.

11. Lee MS, Olson MA. 2010 Protein folding simulations combining self-guided Langevin dynamics and temperature-based replica exchange. Journal of Chemical Theory and Computation 6: 2477–2487.

12. Zwanzig RW. 1954 High temperature equation of state by a perturbation method. I. Nonpolar gases. Journal of Chemical Physics 22: 1420–1426.

13. Straatsma TP, McCammon JA. 1992 Computational alchemy. Annual Review of Physical Chemistry 43: 407–435.

14. Tidor B, Karplus M. 1991 Simulation analysis of the stability mutant R96H of T4 lysozyme. Biochemistry 30: 3217-28.

15. Worth CL, Preissner R, Blundell, TL. 2011 SDM--a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Research 39 (Web Server issue), W215–222.

16. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. 2005 The FoldX web server: an online force field. Nucleic Acids Research 33(Web Server issue), W382–388.

17. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. 2011 PoPMuSiC 2.1: A web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics. 12: 151. doi: 10.1186/1471-2105-12-151.

18. Tian J, Wu N, Chu X, Fan Y. 2010 Predicting changes in protein thermostability brought about by single- or multi-site mutations. BMC Bioinformatics. 11: 370. doi: 10.1186/1471-2105-11-370.

19. Jia L, Yarlagadda R, Reed CC. 2015 Structure based thermostability prediction models for protein single point mutations with machine learning tools. PLoS One 10: e0138022. doi: 10.1371.

20. Khan S, Vihinen M. 2010 Performance of protein stability predictors. Human Mutation 31: 675–684.

21. Olson MA, Zabetakis D, Legler PM, Turner KB, Anderson GP, Goldman ER. 2015 Can template-based protein models guide the design of sequence fitness for enhanced thermal stability of single domain antibodies? Protein Engineering Design & Selection 28: 395-402.

22. Lauck F, Smith CA, Friedland GF, Humphris EL, Kortemme T. 2010 RosettaBackrub--a web server for flexible backbone protein structure modeling and design. Nucleic Acids Research 38(Web Server issue): W569–575.

23. Smith CA, Kortemme T. 2011 Predicting the tolerated sequences for proteins and protein interfaces using RosettaBackrub flexible backbone design. PLoS One 6: e20451. doi: 10.1371/journal.pone.0020451.

24. Barua B, Lin JC, Williams VD, Kummler P, Neidigh JW, Andersen NH. 2008 The Trp-cage: optimizing the stability of a globular miniprotein. Protein Engineering Design & Selection 21: 171–185.

25. RosettaCommons. 2010 Available at: http://www.rosettacommons.org. Accessed 2015 December 22.

26. Zhou H, Zhou Y. 2002 Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science 11: 2714–2726.

27. Wu X, Brook BR. 2003 Self-guided Langevin dynamics simulation method. Chem Phys Letters 381: 512–518.

28. Mackerell AD Jr, Feig M, Brooks CL 3rd. 2004 Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem 25: 1400–15.

29. Lee MS, Feig M, Salsbury FR Jr, Brooks CL 3rd. 2003 New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. J Comput Chem 24: 1348–56.

30. Olson MA, Lee MS. 2013 Structure refinement of protein model decoys requires accurate side-chain placement. Proteins 81: 469–78.

31. Ryckaert J-P, Ciccotti G, Berendsen HJC. 1977 Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phy 23: 327–41.

32. Feig M, Karanicolas J, Brooks CL 3rd. 2004 MMTSB Tool Set: Enhanced sampling and multiscale modeling methods for applications in structural biology. J Mol Graph Model 22: 377–95.

33. Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, et al. 2009 CHARMM: The biomolecular simulation program. J Comput Chem 30: 1545–1614.

34. Holm L, Sander C. 1993 Protein structure comparison by alignment of distance matrices. J Mol Biol 233: 123-38.

35. Krivov GG, Shapovalov MV, Dunbrack RL Jr. 2009 Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77: 778–95.

36. Warshel A, Papazyan A. 1998 Electrostatic effects in macromolecules: fundamental concepts and practical modeling. Curr Opin Struct Biol 8: 211-7.

37. Sham YY, Muegge I, Warshel A. 1998 The effect of protein relaxation on charge-charge interactions and dielectric constants of proteins. Biophys J 74: 1744–1753.

38. Olson MA. 1998 Mean-field analysis of protein-protein interactions. Biophys Chem 91: 219-229.

39. Olson MA, Reinke LT. 2000 Modeling implicit reorganization in continuum descriptions of protein-protein interactions. Proteins 38: 115–119.

40. Li L, Li C, Zhang Z, Alexov E. 2013 On the dielectric "constant" of proteins: smooth dielectric function for macromolecular modeling and its implementation in DelPhi. J Chem Theory Comput 9: 2126-2136.

41. Cumberworth A, Bui JM, Gsponer J. 2015 Free energies of solvation in the context of protein folding: Implications for implicit and explicit solvent models. J Comput Chem doi: 10.1002/jcc.24235.

Table 1. Measured T_m and K_D for wild-type A3 and mutants.

A3 Clone	T_m (°C)	K_D (nM)	Structural Region
WT	85	0.23	
S25I	85	0.19	CDR1
F29A	84	0.20	CDR1
F29L	85	0.18	CDR1
P55S	83	3.30	CDR2
Y59A	82	0.45	CDR2
Y59P	84	No binding	CDR2
R70A	74	0.33	FR3
S74A	83	0.21	FR3
A75R	84	0.52	FR3
Y98A	75	0.34	FR3
D102A	84	0.09	CDR3
D102R	73	0.29	CDR3
K104G	78	0.21	CDR3
M111A	76	0.25	CDR3
M111T	76	0.50	CDR3
M111A/V107I	76	0.60	CDR3
V116A	80	0.14	CDR3
V116Y	79	0.13	CDR3

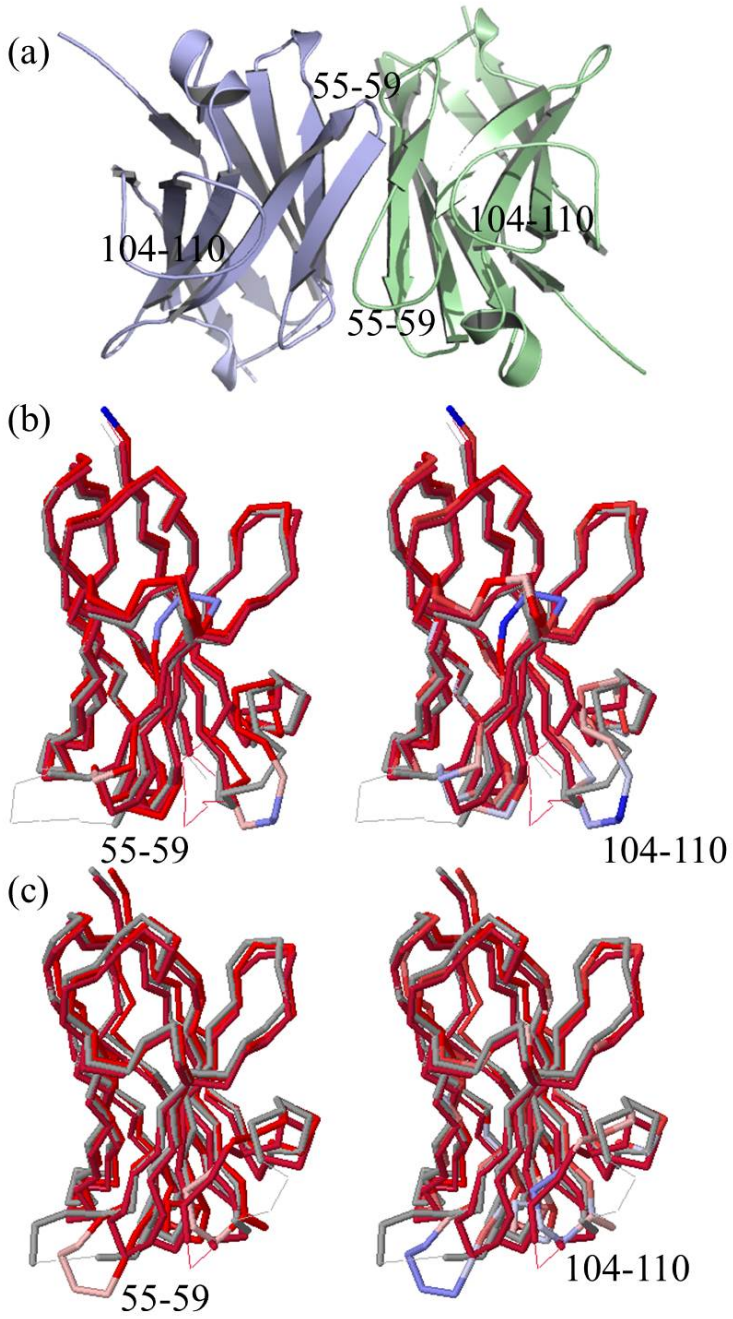

Figure 1. X-ray crystallographic structure of the asymmetric homodimeric A3 and structural neighbors from a Dali search of the single-chain conformers of the assembly. (a) Conformers of A3 where the color green denotes the A chain and blue color the B chain. (b) Structure-structure alignment of the A-chain with sdAbs 3stb and 1i3v, where on the left shows structural differences and the right sequence differences. The color spectrum runs from red to blue, where red designates structural and/or sequence conservation, while blue show divergence. (c) Alignments of the B-chain with sdAbs 3stb and 1i3v.

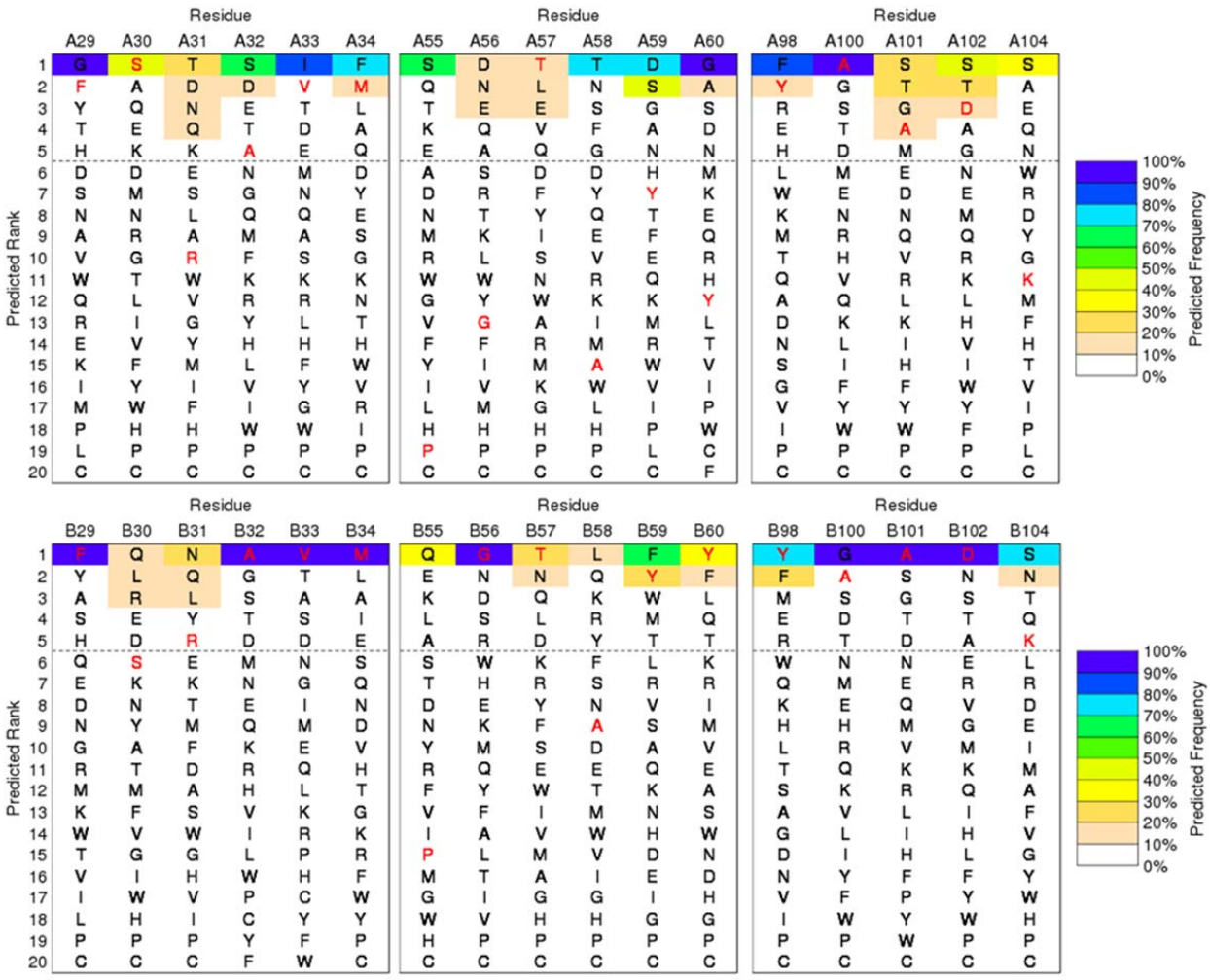

Figure 2. Selected sequence-tolerance profiles of the A-chain conformer (top graph) and B-chain conformer (bottom graph) determined by the RosettaBackrub method. Sequences colored red denote the wild-type residue. The color spectrum of each profile designates the predicted frequency of sequence-site placement along the A- or B-chain conformers. The horizontal line is arbitrarily positioned to highlight the five top-ranking sequences.

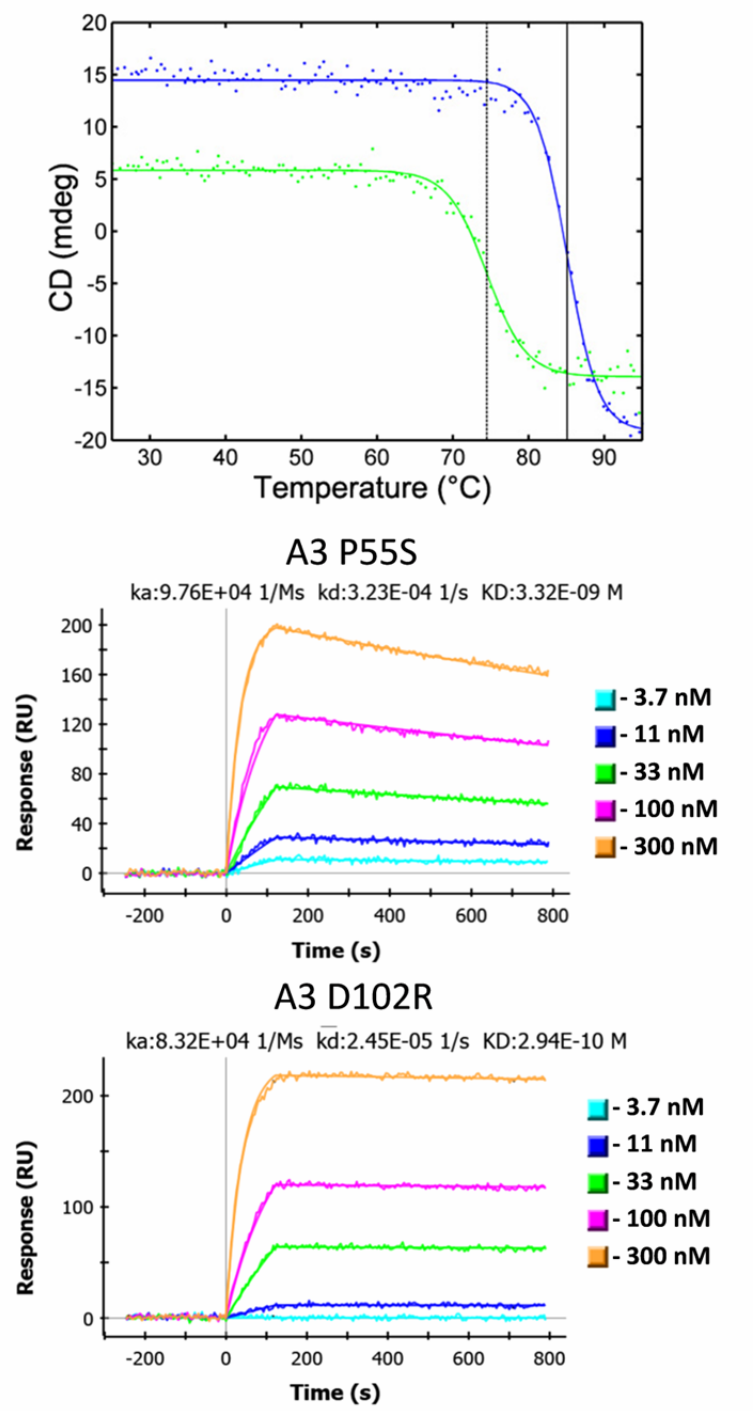

Figure 3. Representative experimental data. Top graph shows denaturation curves of A3 mutants F29L (blue) and Y98A (green). Vertical lines have been drawn through the inflection points, dotted for Y98A and solid for F29L. Middle and lower graphs are SPR data for mutants P55S and D102R. The on rate, off rate and calculated dissociation constant (K_D) of A3 binding the target SEB are shown above the data for these mutants.

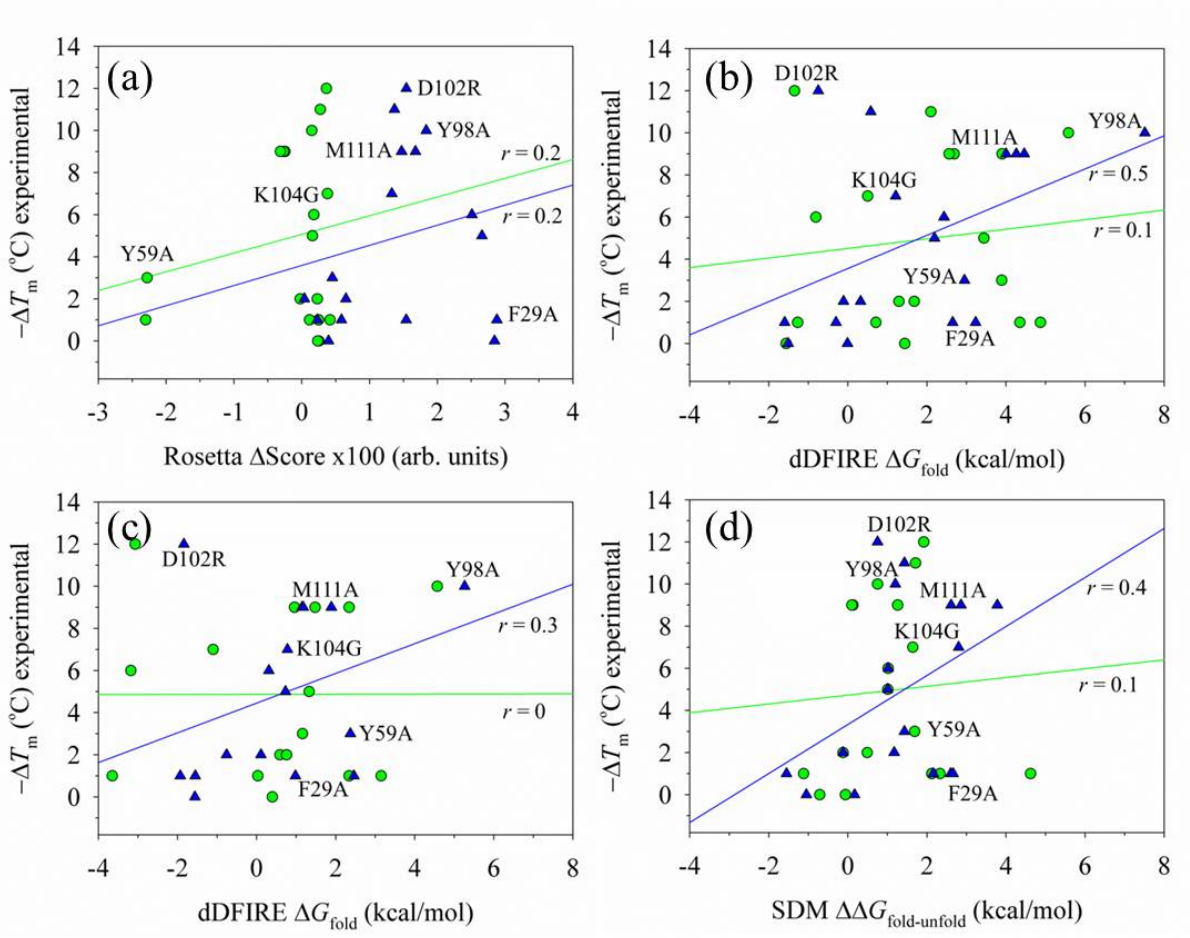

Figure 4. Scatter plots of scoring the effect of 18 residue substitutions relative to the wild-type structure and their relationship to the experimentally determined changes in melting temperatures (ΔT_m). To account for differences in physical units between the computed ΔG and the experimental ΔT_m, the free-energy coordinate can be arbitrarily scaled by a "reference-state." For illustrations (**a**)-(**c**), conformations were generated by the RosettaBackrub simulation method starting with either the crystallographic A-chain conformer (denoted as green-colored circles) or the B-chain conformer (blue triangles). The first two plots were computed as statistical averages over each modeled conformational ensemble. (**a**) Application of the Rosetta 3.1 scoring function to the mutations, yielding linear regression correlation coefficients of $r = 0.2$ for both chains. (**b**) Scoring of the generated ensemble by dDFIRE, yielding $r = 0.1$ for the A-chain and $r = 0.5$ for the B-chain. (**c**) The fist-order rank conformer determined by dDFIRE scoring for the two chains ($r \sim 0$ for the A-chain and $r = 0.3$ for the B-chain). (**d**) Application of the SDM modeling method using the starting crystallographic A-chain and B-chain conformers to calculate changes in free-energies and their relationship to T_m ($r \sim 0$ for A-chain and $r = 0.3$ for the B-chain).

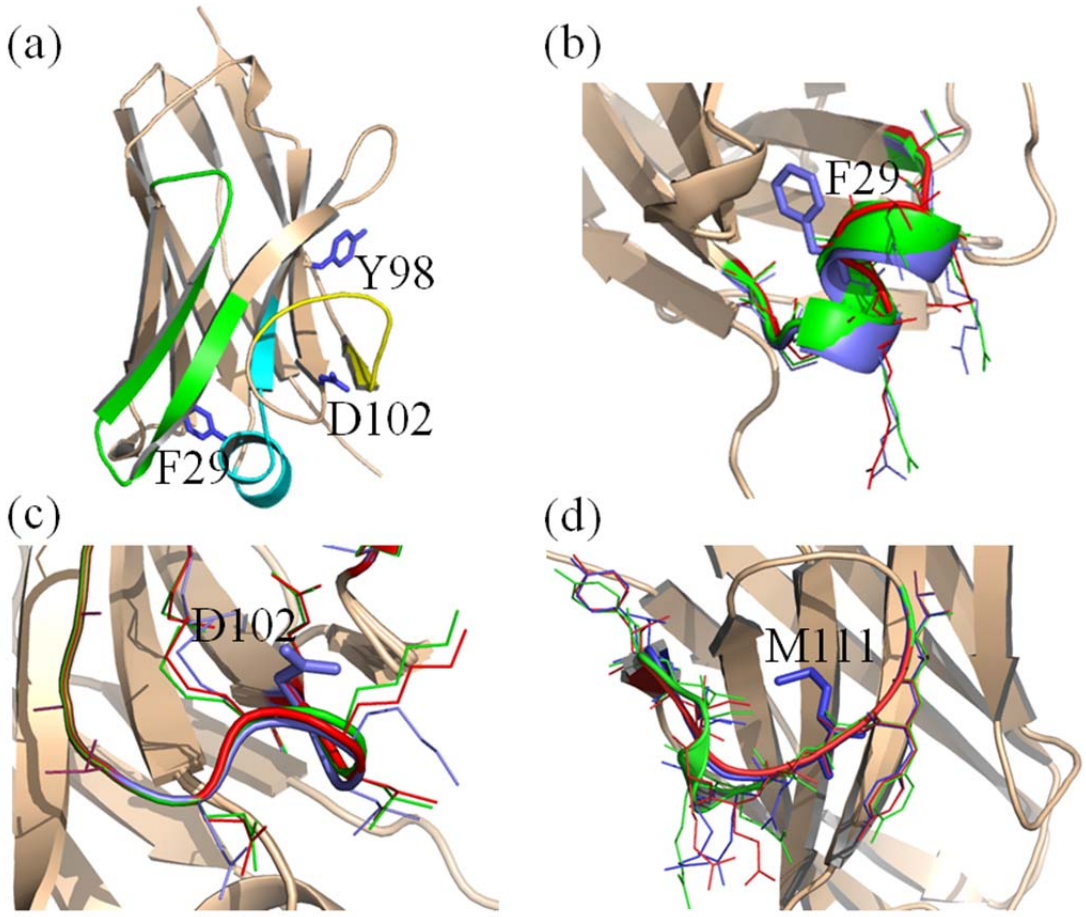

Figure 5. Molecular illustrations of the modeled folded structure of the wild-type sdAb conformation and selected mutants determined by RosettaBackrub simulations. The structures are (a) wild-type B-chain conformer showing CDR1 (colored cyan), CDR2 (green), CDR3 (yellow) and several mutational sties; (b) mutant F29A; (c) D102R; and (d) M111A. For figures (b-d), the wild-type conformation surrounding a selected sequence position is shown in blue, the top-rank ordered dDFIRE structure from the generated ensemble is shown in green and the lowest-rank ordered dDFIRE conformation is shown in red. Surrounding local side-chains are highlighted in the respective colors.

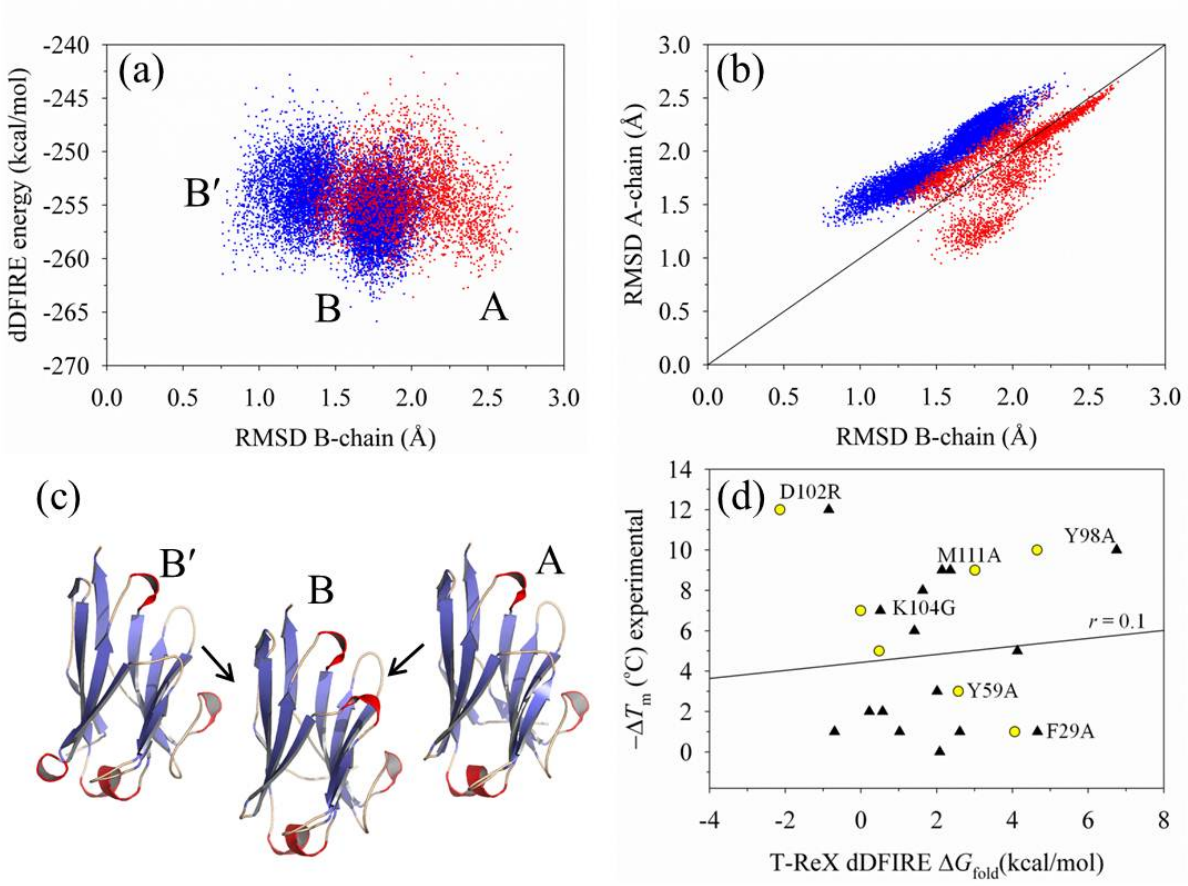

Figure 6. Conformations generated by the T-ReX simulation of A3. (**a**) dDFIRE energy landscape as a function of C_α RMSD computed for each conformation relative to the starting B-chain. The color blue designates conformers among the final ensemble that originated from the starting B-chain and red denotes structures that originated from the starting A-chain conformation. (**b**) Scatter plot of C_α RMSD computed for each conformation against the starting A-chain and B-chain structures. (**c**) Selected conformations (from left to right) that representative a conformer from the large basin of low-RMSD states, the overall top-rank dDFIRE scoring conformation that originated from the B-chain, and the top-rank conformer derived from the A-chain. (**d**) RosettaBackrub simulations using the starting conformation that corresponds to the top-rank B-chain (denoted as black-colored triangles) and, for a selected subset of amino acid substitutions, predictions using five conformers taken from the low-RMSD basin to compute an average ΔG_{fold} (circles colored yellow).

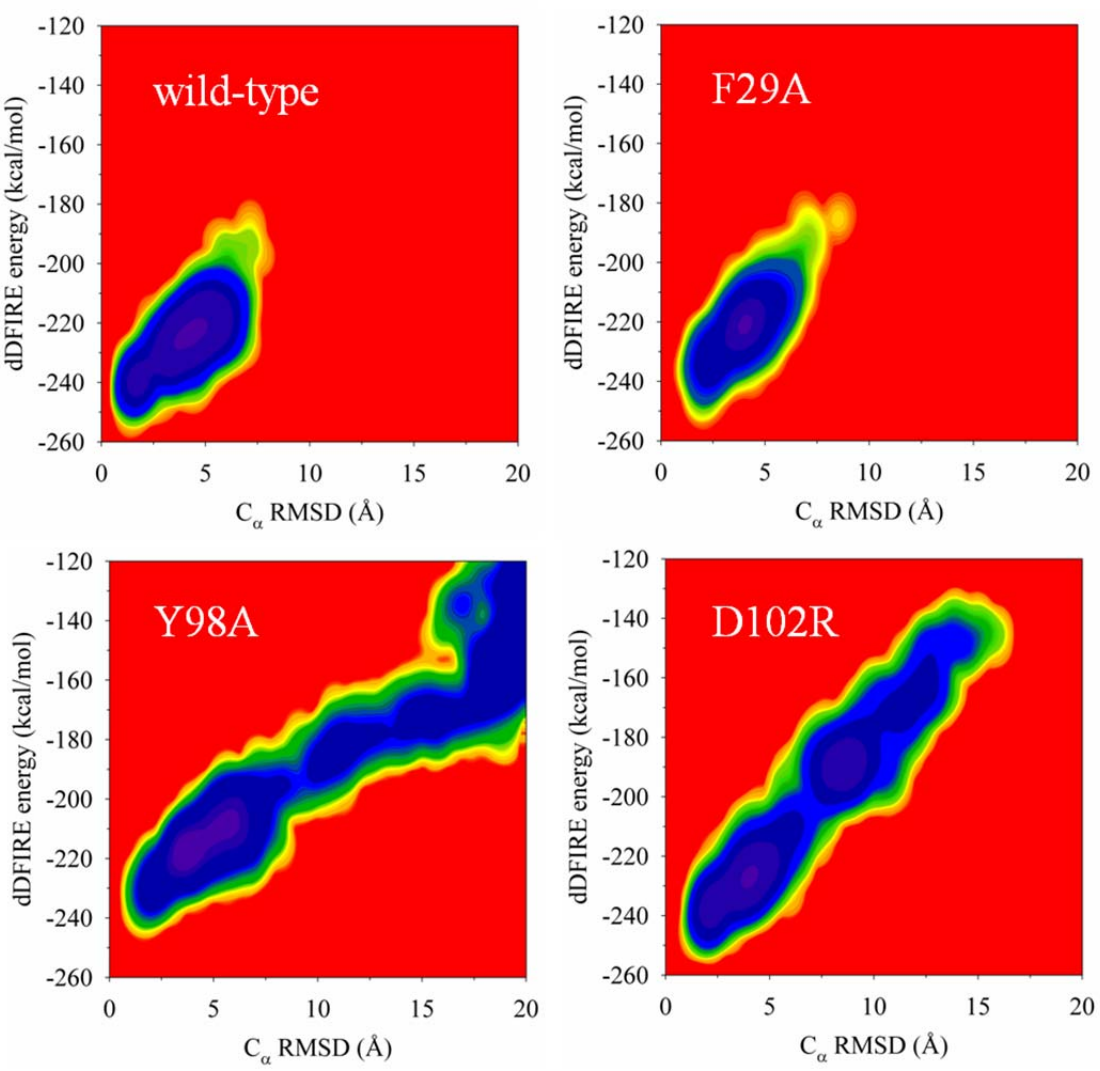

Figure 7. Two-dimensional population density profiles as a function of the statistical potential dDFIRE and Cα RMSD from the starting crystallographic conformer B-chain at a temperature where the mutant Y98A is modeled to exhibit equal population between native and non-native states. Native-like states are defined as displacements from the starting structures by the approximation Cα-RMSD < 5 Å. Comparisons are shown for the wild-type B-chain, F29A, Y98A and D102R. The color spectrum contains the extremes of blue for high population density and red for low density.

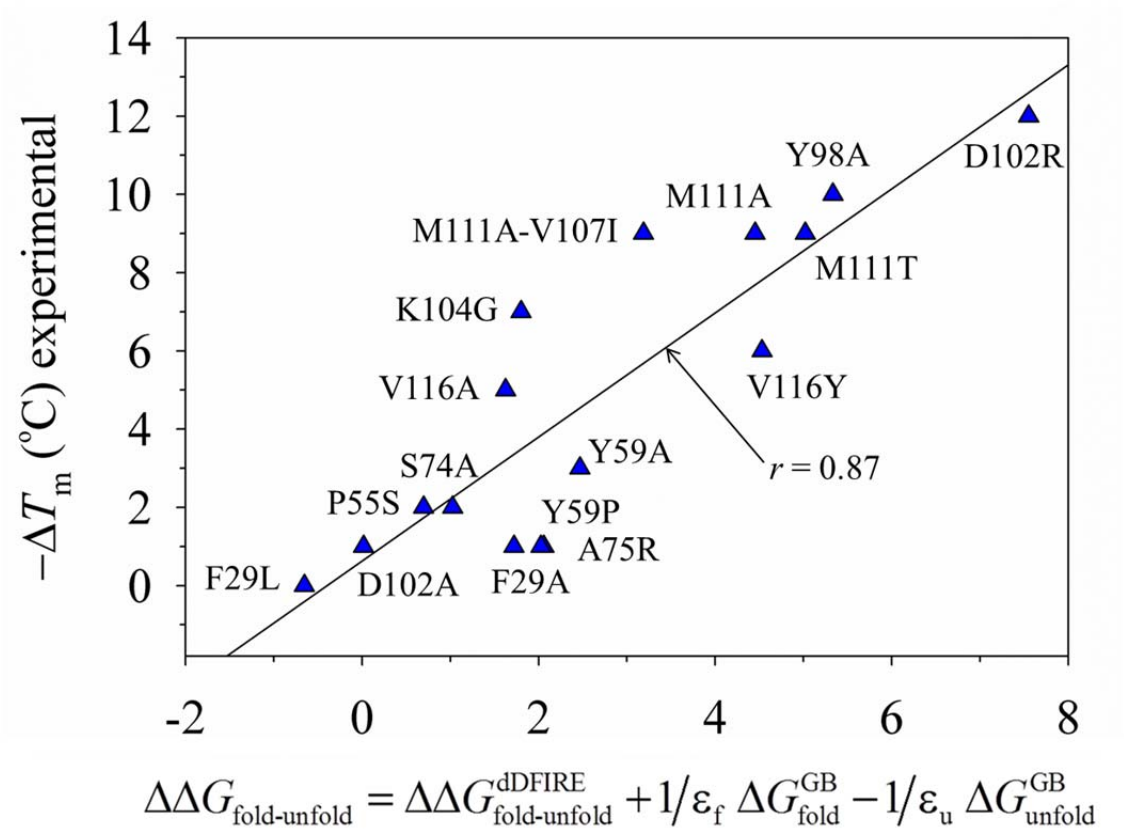

Figure 8. Linear-scaling approximation of scoring 16 sequence substitutions of the B-chain conformer by combing the dDFIRE statistical potential and the GB implicit solvent model applied to a two-state model of the folded and unfolded conformations. The GB solvent term contains scaling parameters that effectively treat reorganization of the modeled conformations. The linear regression correlation coefficient is $r \sim 0.9$.